Employee Performance Analysis

A Project Report

Submitted by :-

Manish Bansal
(Mail id:- man.bansal60@gmail.com)

In partial fulfillment of the requirements

for the award of the certificate of

'Certified Data Scientist'

ACKNOWLEDGEMENT

I express my deep sense of gratitude to our respected and learned guide, ‘Mr. Ashok Kumar A’ for his valuable traning and guidance, we are also thankful to his valuable encouragement he has given us in completing the project.

I am also thankful to all other faculty and staff members of ‘DataMites™ Solutions Pvt Ltd’ for their kind co-operation and help.

Table of Contents

1. Project Summary

INX Future Inc, is one of the leading data analytics and automation solutions provider with over 15 years of global business presence. In recent years, the employee performance indexes are not healthy and this has become a growing concern among the top management. The CEO Mr. Brain, decided to initiate a data science project, which analyzes the current employee data and find the core underlying causes of the performance issues. He also expects a clear indicator of non-performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals. The following insights are expected from this project:

- Department wise performances.
- Top 3 Important Factors effecting employee performance.
- A trained model which can predict the employee performance based on factors as inputs.
- Recommendations to improve the employee performance based on insights from analysis

1.1 Requirement

To analyze and solve the requirement problem statement provided by
IABAC team and we have used the dataset from the below URL.

http://data.iabac.org/exam/p2/data/INX_Future_Inc_Employee_Performance_CDS_Project2_Data_V1.8 .xls

We have processed the dataset for EDA, converting categorical to dummy variables for building a ML model as well as doing scaling techniques can be found under the project submission folder in the file <train_model.ipynb>

1.2 Analysis

- The dataset used is supervised and categorical. Some independent variables are ordinal and a few among them are nominal as well as target variable 'Performance Rating' is ordinal.

- To analyze the dataset, various data processing techniques like Label Encoding and Standardization is used, along that Correlation Coefficient is used to interpret the relationship between variables. The most important features selected are EmpDepartment', 'EmpJobRole', 'EmpEnvironmentSatisfaction','EmpJobLevel', 'EmpLastSalaryHikePercent', 'EmpWorkLifeBalance','ExperienceYearsAtThisCompany', 'ExperienceYearsInCurrentRole','YearsSinceLastPromotion', 'YearsWithCurrManager

- For training the data and predicting the target, algorithms used are Logistic Regression, Decision Tree, Random Forest, XGBoost Classifier and K-Nearest Neighbor

- A separate analysis of Department wise Performance is carried out. Same to find out top 3 Important Factors affecting employee performance.

1.3 Summary

The project was done with the purpose of solving the analyses the current employee data and find the core underlying causes of this performance issues and finding out factors which affected the Performance of the employees, training a model which accurately predicts the Performance Rating of the employee, analyzing the data to provide recommendations to improve the performance and gain insights from the analysis. The following steps were carried out:

- Dataset was import from the URL provided, find out the predictor & target variables and look for missing values. Analysis of Department wise performance as per the problem statement.
- Label Encoding was done for the ordinal columns to make it compatible for Model.
- Calculate correlation coefficient to find out the relationship between variables and then select the important features for analysis.
- Standardizing the data to treat the outliers if any and splitting it into test and train.
- Training the data using algorithms like Logistic Regression, Decision Tree, Random Forest, XGBoost Classifier and K-Nearest Neighbor and checking the accuracy to find out which algorithm is the best.
- Analyzing the model that have the highest accuracy and exporting the model for the new data.

2. Data

2.1 External/Raw Data:

The employee performance date of INX Future Inc. can be downloads from below link:
http://data.iabac.org/exam/p2/data/INX_Future_Inc_Employee_Performance_CDS_Project2_Data_V1.8 .xls

2.2 Processed Data:

Final processed dataset after applying Data cleaning, EDA and other scaling techniques can be found under the project submission folder:
O/P File: <**datasets/ dataset_for_model.csv** >

3. Source code

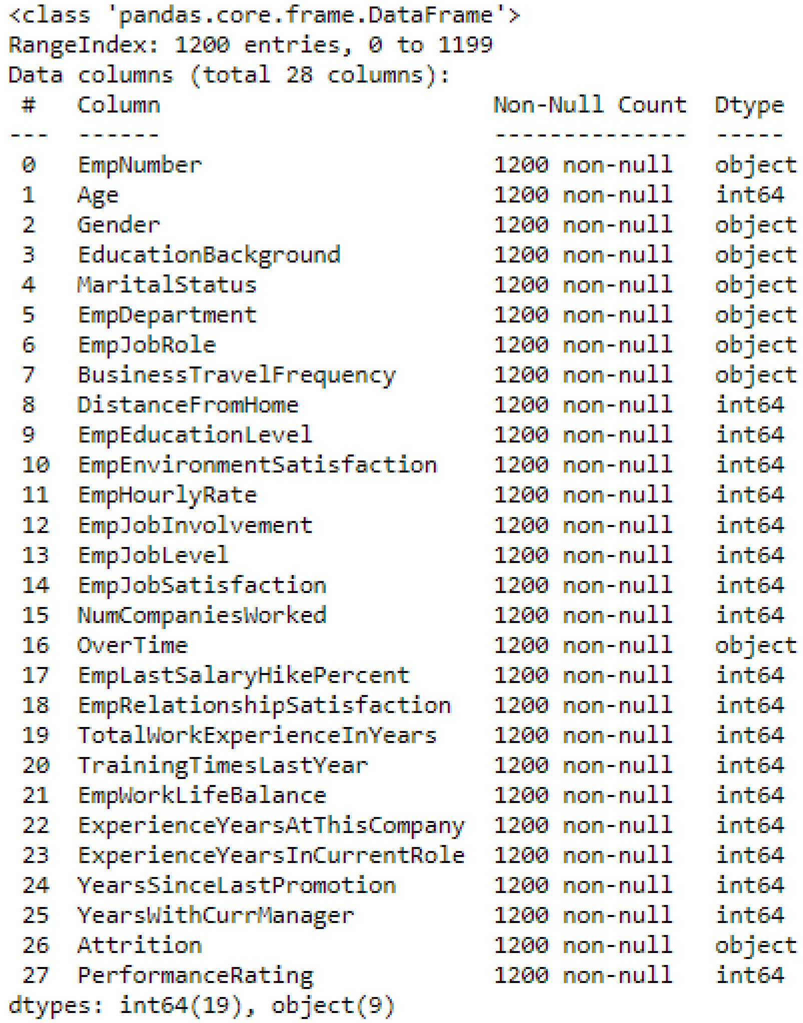
This section will discuss about the Source code used in this project to solve the problem statements as well as to build the model, Data processing, data mugging, exploratory Analysis.

3.1 Data Processing

We will describe the outputs or insights we have extracted from the data_processing.ipynb and data_exploratory_analysis.ipynb.

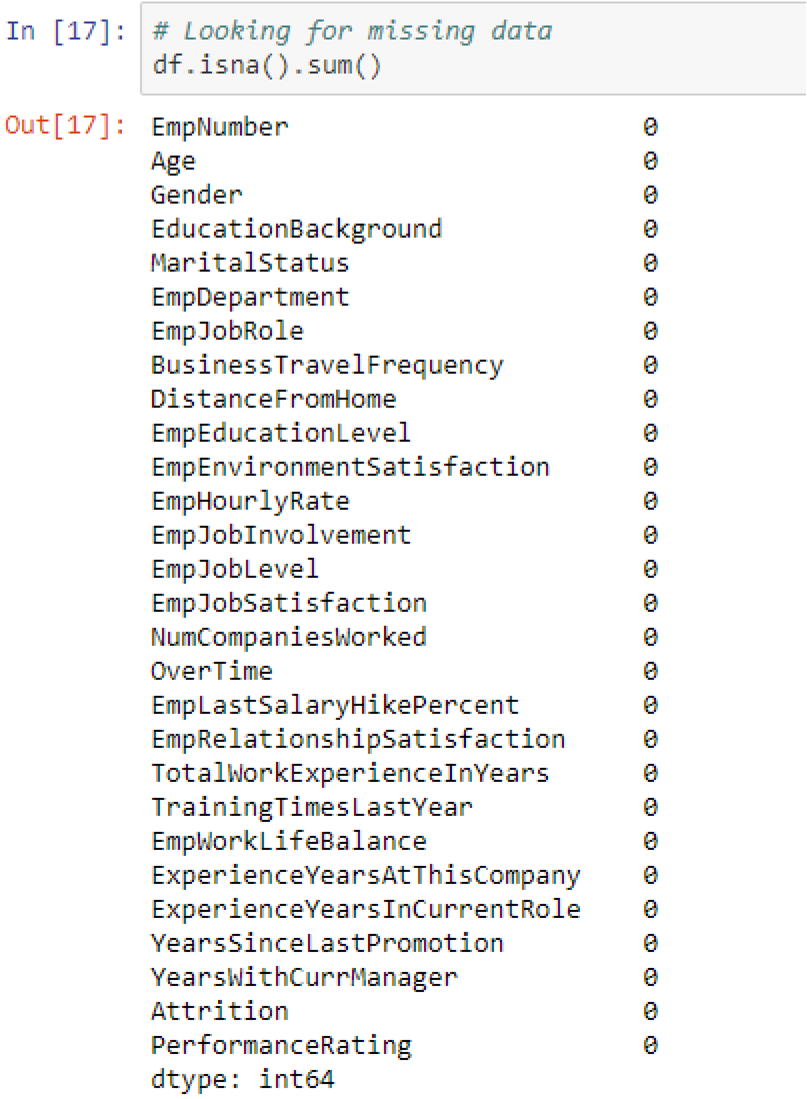
3.1.1 Data_processing.ipynb

- The dataset consists of 28 independent variables (inputs) with dependent variable(output) i.e., PerformanceRating.
- By seeing the shape of data frame, we can say there are total 1200 rows.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 28 columns):
 #   Column                        Non-Null Count  Dtype 
---  ------                        --------------  ----- 
 0   EmpNumber                     1200 non-null   object
 1   Age                           1200 non-null   int64 
 2   Gender                        1200 non-null   object
 3   EducationBackground           1200 non-null   object
 4   MaritalStatus                 1200 non-null   object
 5   EmpDepartment                 1200 non-null   object
 6   EmpJobRole                    1200 non-null   object
 7   BusinessTravelFrequency       1200 non-null   object
 8   DistanceFromHome              1200 non-null   int64 
 9   EmpEducationLevel             1200 non-null   int64 
 10  EmpEnvironmentSatisfaction    1200 non-null   int64 
 11  EmpHourlyRate                 1200 non-null   int64 
 12  EmpJobInvolvement             1200 non-null   int64 
 13  EmpJobLevel                   1200 non-null   int64 
 14  EmpJobSatisfaction            1200 non-null   int64 
 15  NumCompaniesWorked            1200 non-null   int64 
 16  OverTime                      1200 non-null   object
 17  EmpLastSalaryHikePercent      1200 non-null   int64 
 18  EmpRelationshipSatisfaction   1200 non-null   int64 
 19  TotalWorkExperienceInYears    1200 non-null   int64 
 20  TrainingTimesLastYear         1200 non-null   int64 
 21  EmpWorkLifeBalance            1200 non-null   int64 
 22  ExperienceYearsAtThisCompany  1200 non-null   int64 
 23  ExperienceYearsInCurrentRole  1200 non-null   int64 
 24  YearsSinceLastPromotion       1200 non-null   int64 
 25  YearsWithCurrManager          1200 non-null   int64 
 26  Attrition                     1200 non-null   object
 27  PerformanceRating             1200 non-null   int64 
dtypes: int64(19), object(9)
```

- There is no missing value observed in the dataset.



```
In [17]: # Looking for missing data
         df.isna().sum()
```

```
Out[17]: EmpNumber                       0
         Age                             0
         Gender                          0
         EducationBackground             0
         MaritalStatus                   0
         EmpDepartment                   0
         EmpJobRole                      0
         BusinessTravelFrequency         0
         DistanceFromHome                0
         EmpEducationLevel               0
         EmpEnvironmentSatisfaction      0
         EmpHourlyRate                   0
         EmpJobInvolvement               0
         EmpJobLevel                     0
         EmpJobSatisfaction              0
         NumCompaniesWorked              0
         OverTime                        0
         EmpLastSalaryHikePercent        0
         EmpRelationshipSatisfaction     0
         TotalWorkExperienceInYears      0
         TrainingTimesLastYear           0
         EmpWorkLifeBalance              0
         ExperienceYearsAtThisCompany    0
         ExperienceYearsInCurrentRole    0
         YearsSinceLastPromotion         0
         YearsWithCurrManager            0
         Attrition                       0
         PerformanceRating               0
         dtype: int64
```

3.1.2 Data_exploratory_analysis.ipynb

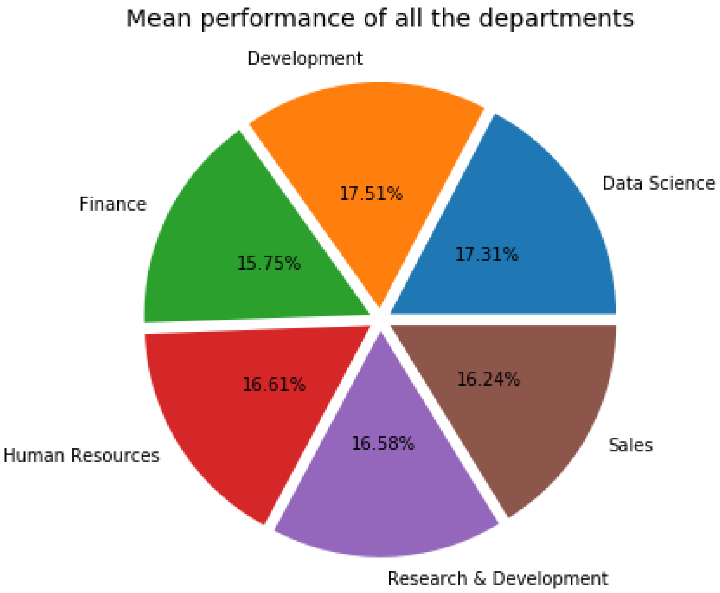
Under EDA, we have studied the below components:

- Understanding the variables
- Analyzing relationships between variables
- Data Munging
- Analysis of Department wise Performance
- Finding the top 3 Important Factors effecting employee performance

3.1.2.1 Problem statement 1: Analysis of Department wise Performance

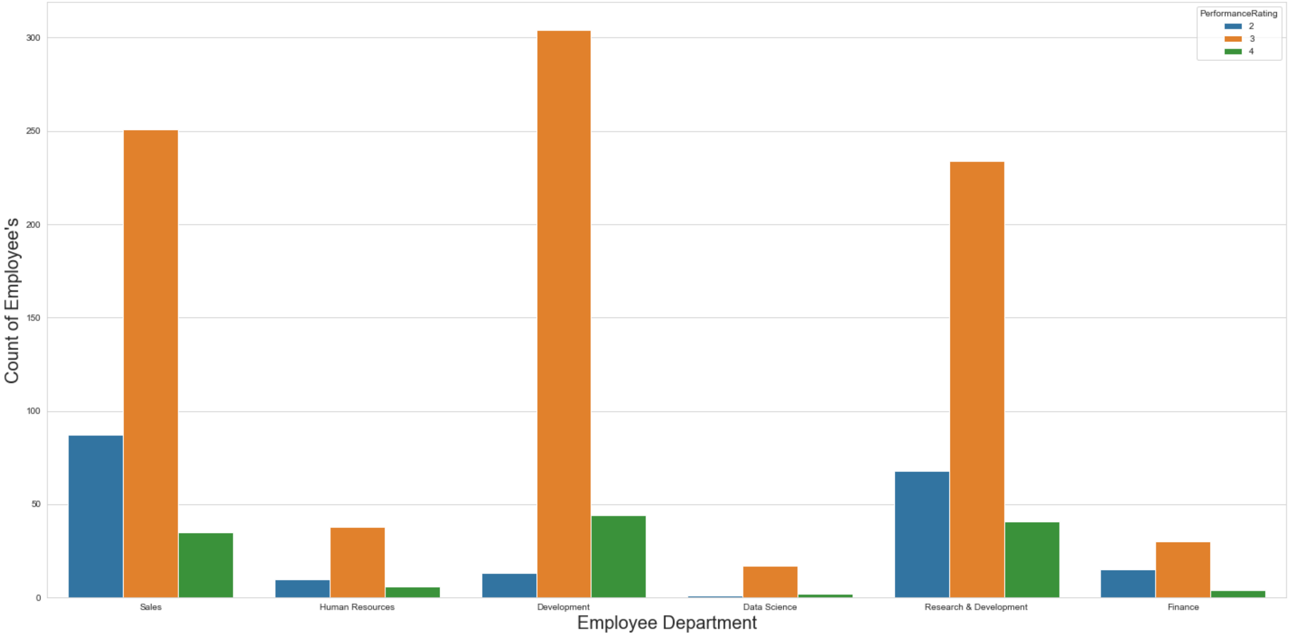
'**Pie plot Insights**'

- '**HIGHEST**' Performance percentage is of Development Department i.e., 17.51%
- '**LOWEST**' Performance percentage is of Finance Department i.e., 15.75%

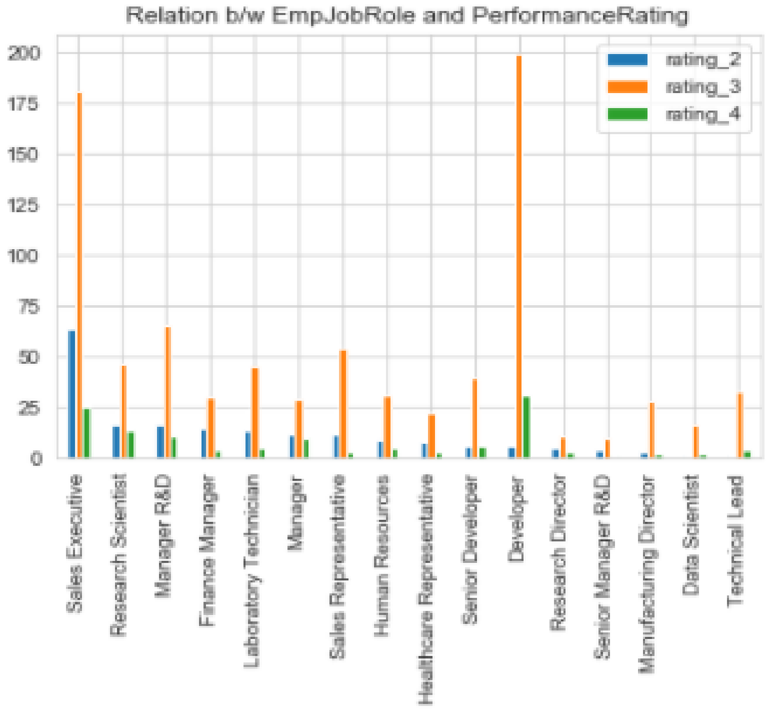


'**Bar Plot Insights**'

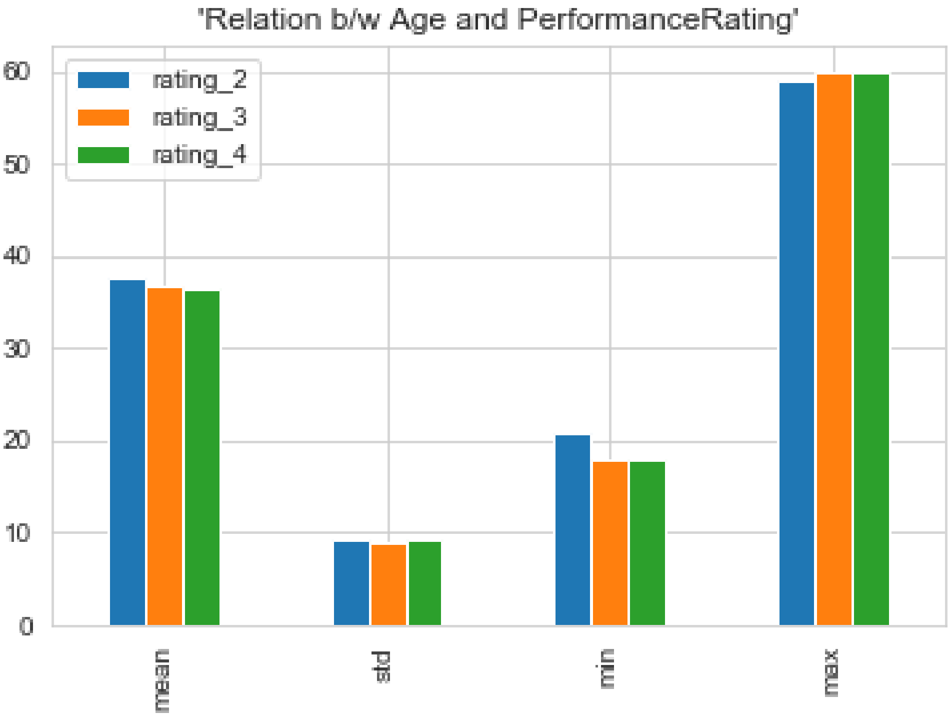
- Employees provided with PerformanceRating '**Performance Score-2**' is highest in Sales, Research & Development departments.
- Employees provided with PerformanceRating '**Performance Score-3**' is highest in Development department.
- Employees provided with PerformanceRating '**Performance Score-4**' is highest in Development department.



- Employee who are less involved or intermediately involved in their job’s role have performance rating more towards 3 and 2 whereas sales and finance & developers seems to enjoy their jobs.

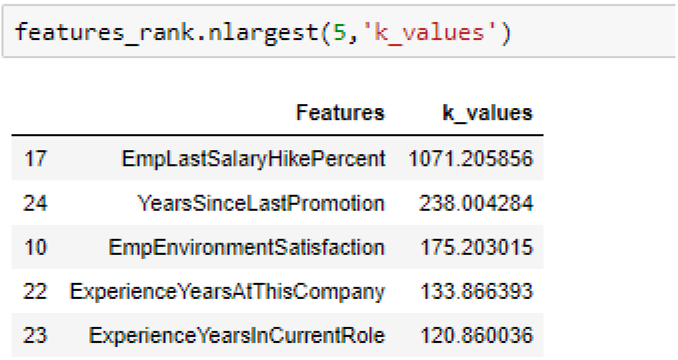


- Rating 4 employees have less age and very active in their work then the higher-level employees



3.1.2.2 Problem statement 2: Top 3 Important Factors effecting employee performance

- By using SelectKBest Algorithm we have find the top 5 independent variables effecting the employee's performance.

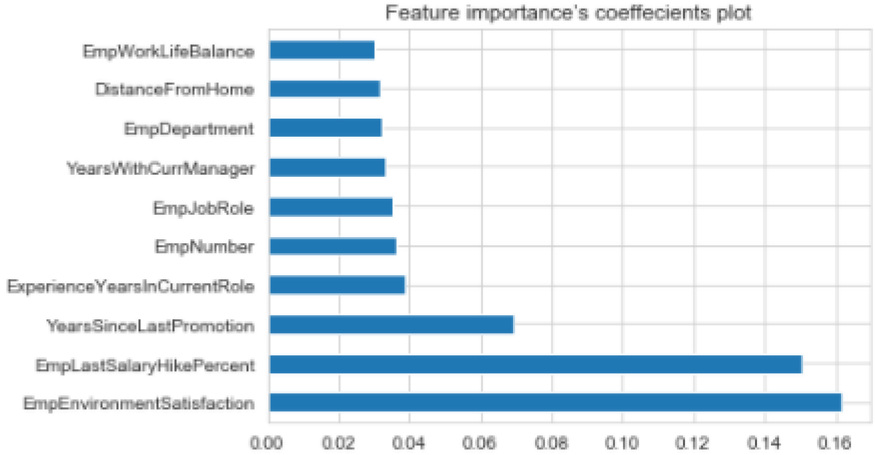


```
features_rank.nlargest(5,'k_values')
```

	Features	k_values
17	EmpLastSalaryHikePercent	1071.205856
24	YearsSinceLastPromotion	238.004284
10	EmpEnvironmentSatisfaction	175.203015
22	ExperienceYearsAtThisCompany	133.866393
23	ExperienceYearsInCurrentRole	120.860036

- From below plot of Feature importance's coefficients we can see the mentioned features have high coefficient value

1. EmpLastSalaryHikePercent
2. YearsSinceLastPromotion
3. EmpEnvironmentSatisfaction
4. ExperienceYearsAtThisCompany
5. ExperienceYearsInCurrentRole



- In addition to the above results captured based on the 'SelectKBest Algorithm' and 'Feature Importance' techniques.

  Correlation values for the Top-5 features are as below:

  1. EmpLastSalaryHikePercent : 0.333722
  2. YearsSinceLastPromotion : -0.167629
  3. EmpEnvironmentSatisfaction : 0.395561
  4. ExperienceYearsAtThisCompany : -0.111645
  5. ExperienceYearsInCurrentRole : -0.147638

So, based on the above values as well as in addition to the above results captured based on the 'SelectKBest Algorithm' and 'Feature Importance' techniques, we can conclude that below are the TOP 3 features that will affect the Employee Performance rating criteria are
1. Employee EnvironmentSatisfaction
2. Employee Last Salary Hike Percent
3. Years Since Last Promotion

3.2 Models

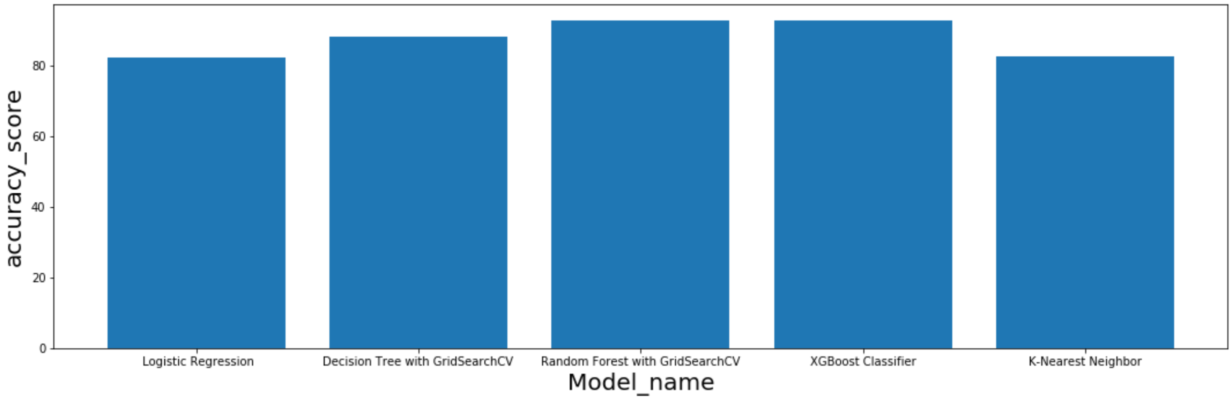
3.2.1 Train_model.ipynb

Some insights we got from the correlation of date set which helped in modeling the model

- The features that are positively correlated are Environment Satisfaction, Last Salary Hike Percent & Work life Balance. This means that if these factors increase, Performance Rating will increase. On the other hand, the features that are negatively correlated are Years Since Last Promotion, Experience Years at this Company, Experience years in Current Role & Years with Current Manager. This means that if these factors increase, Performance Rating will go down.

The model source code can be look into using the **train_model.ipynb**

1. For modeling the data, first we have Standardize the data (method to treat outliers) and then it is splitted into train and test.
2. Data is trained using algorithms like Logistic Regression, Decision Tree, Random Forest, XGBoost Classifier and K-Nearest Neighbor. By looking the below graph we have compared the accuracy to find out which algorithm is the best.



Problem statement 3: A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees.

By analyzing the above graph be figure out that Random Forest with GridSearchCV and XGBoost Classifier gives the maximum and same accuracy of approx. 93%

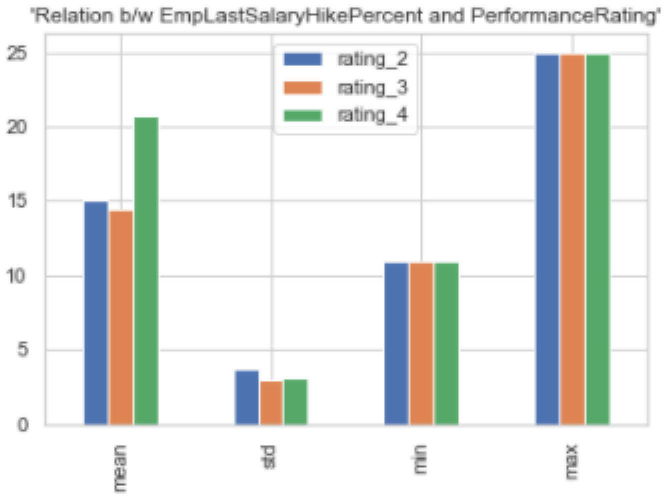
By importing 'INX_Future_Inc.ml' in predict_model.ipynb we can predict the rating of the newly input data.

4 visualization

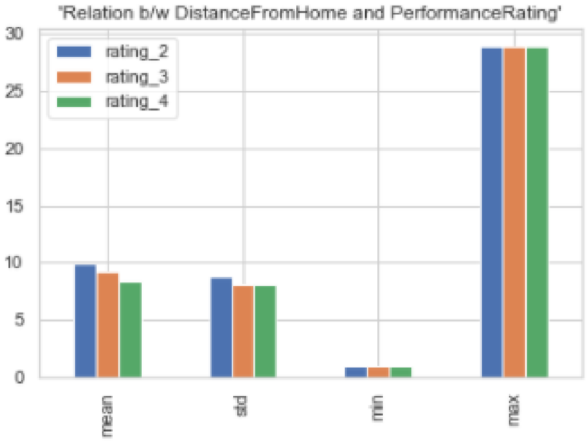
Visualization is done to find some interesting/ extra insights from the data set and source code ca be look into by visualize.ipynb

Some insights are listed below: -

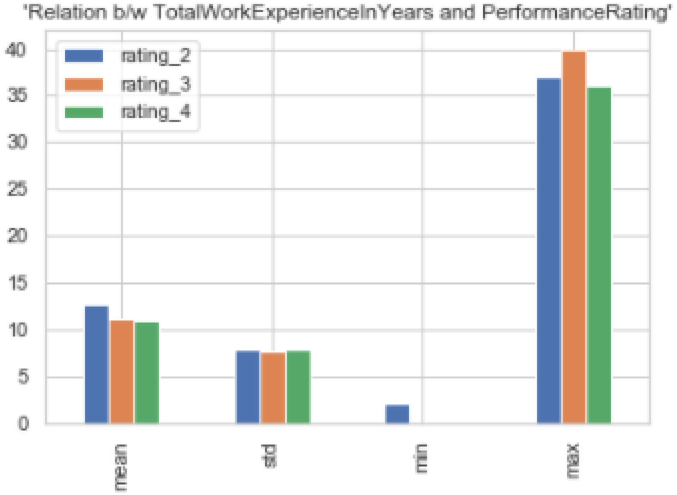
- Employees have Rating-4 are have more salary hike percentage than comparing with others means they are doing good and get the salary hike more.



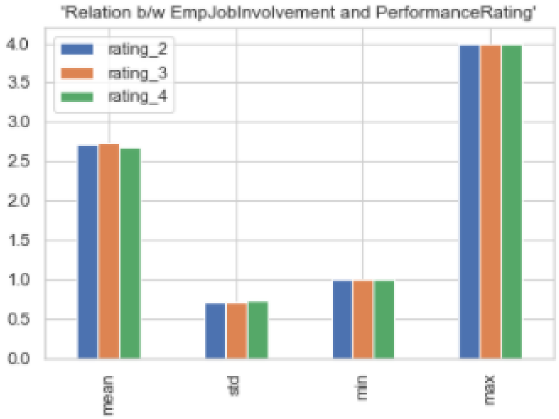
- Rating 4 employees have less distance from their home to office and can reach the office at much early then others in critical work situation



- we can conclude from the above graph that less work experience is very active in learning new skills and reproducing them into their work whereas rating 2 employees are not much good in upgrading themselves.



- All the rating employees are equally involved in their job as the mean values are approximately equal.



5 **Problem statement 4:** Recommendations to improve the employee performance based on insights from analysis

- The company should provide a better environment as it increases the performance drastically. The company should increase the salary of the employee from time to time and help them maintain a work life balance.
- Secondly, there should be quarterly reviews of the employee by employer as well as a monthly meeting with manager can help the employees to follow up as they will get to know where they are lacking and which skill set, they need help which can improve the quality of work also it will help employee to develop himself.
- Taking consideration of quarterly appraisal can help the company to track constantly the performance, with that it will help employer to take the decision to what extend they can hike the salary.
- Addition to that, managers or leads can set the performance goals.

6 REFERENCES

[1] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[2] https://matplotlib.org/3.3.3/contents.html

[3] https://seaborn.pydata.org/

[4] https://pandas.pydata.org/docs/